Chapter 11 Statistics – (Study) Formula Sheet

11.1 – Statistical Studies

Sampling techniques

equal chance

select every 5th

Easiest (Biased)

① split by characteristic
② + Randomly sample within each group (strata)

① mini-populations
② + census each randomly selected group (cluster)

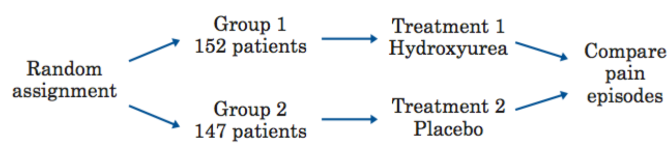
Observational Study vs Experiment

- **Observational study** – Observes existing data.
  - Can reveal association or correlation between variables, but not causation.

- **Experiment** – Generates data to help identify cause-and-effect relationships.
  - Imposes treatments and controls randomly to groups.

Principles of Experimental Design

1. Randomize the control and treatment groups.
2. Control for outside effects on the variable.
3. Replicate the experiment a significant number of times to see meaningful patterns.



11.2 – Displaying Data

Frequency Tables

- Summarize datasets by counting the number of observations for each category, distinct value or interval.

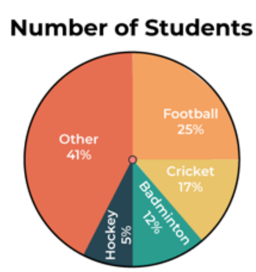
Type of Computer	Frequency	Percent
Desktop	11	11/50 = 22%
Laptop	23	23/50 = 46%
Notebook	9	9/50 = 18%
Tablet	7	7/50 = 14%

Number of Pets	Frequency
1-2	7
3-4	3
5-6	3
7-8	2

Total = 15

Graphical Displays of Data

- Pie charts (categorical data)
  - Compare parts to a whole (slices are proportion of a category).



Examples

a) What percent of observations have between 1 and 4 pets inclusive?

$\frac{7+3}{15} = \frac{10}{15} = \boxed{66.7\%}$

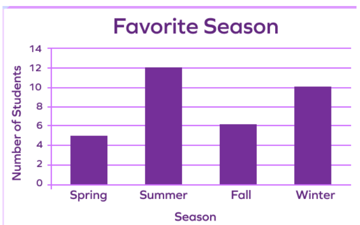
b) What percent of students prefer Football or Hockey?

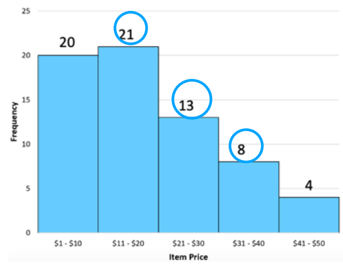
% Football + % Hockey

$= 25\% + 5\%$

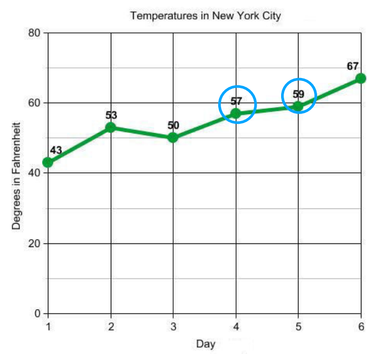
$= \boxed{30\%}$

- Bar graphs (categorical data) and Histograms (numeric data)
    - Height of bar represents amount of data in each category (counts or relative frequencies).





- Line graph
    - Shows changes in a numerical variable over time.



c) Bar graph – Which season has the highest frequency?

Summer → 12

d) Histogram – How many items cost between \$11 and \$40 inclusive?

$21 + 13 + 8 = \boxed{42}$

e) How many days was the temperature between 55 and 60 °F?

$\boxed{2 \text{ days}}$

11.3 – Describing and Analyzing Data

Measures of Center

- **Mean** (average) = $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$
    - NOT resistant → Affected by outliers
- **Median** (middle)
    - The middle value in an ordered list.
    - Resistant → NOT affected by outliers.
- **Mode** (most common)
    - The most frequently occurring value(s).
    - Resistant → NOT affected by outliers.
    - Only measure of center that can be used with categorical data.

Measures of Spread

- **Range** = Max – Min
- **Standard deviation**
    - Measures average distance from the mean.
    - (Don't calculate by hand).

Use calculator to answer these if possible!!!

Example

Dataset: 1, 2, 7, 3, 6, 9, 1, 0, 4, 7

$n = 10$

a) Find the mean.

Calc: 1-Var Stats (Data in L_1)

By hand

$\dfrac{1 + 2 + \cdots + 7}{10} = \boxed{\bar{x} = 4}$

b) Find the median. $\boxed{\text{med} = 3.5}$

0, 1, 1, 2, 3, 4, 6, 7, 7, 9

$(3+4)/2 = 3.5$

c) Find the mode.

1 + 2 ⇒ occur twice

d) Find the range.

Range = max – min

$= 9 - 0 = \boxed{9}$

e) Find the sample standard deviation.

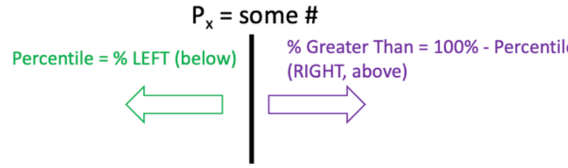
from calc

$\boxed{s_x = 3.091}$

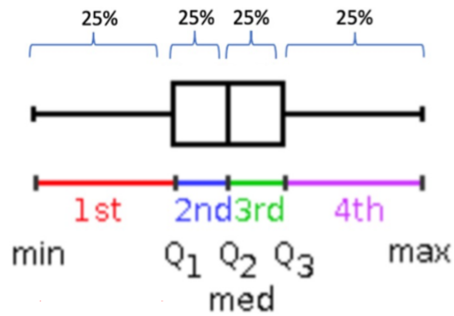
sample

Measures of Relative Position

- A **percentile** tells you the percent of observations/individuals you are higher than.



- **Quartiles** are specific percentiles.
  - Q_1 is the 25th Percentile.
  - Q_3 is the 75th Percentile.
  - Q_2 is the 50th Percentile = Median.



- **Inner Quartile Range (IQR)**
  - $\boxed{IQR = Q_3 - Q_1}$

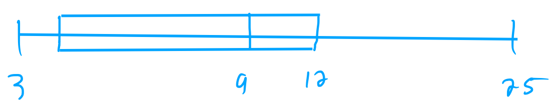
- **5-number summary**
  - Min, Q_1, Med, Q_3, Max → Points of a boxplot

- Example: Calculate the 5-number summary and sketch a boxplot for the following dataset.
  - 12, 3, 4, 7, 21, 3, 9, 8, 10, 11, 25, 11, 13, 4, 5

By hand: ~~3~~, ~~3~~, ~~4~~, 4, ~~5~~, ~~7~~, ~~8~~, 9, ~~10~~, ~~11~~, ~~11~~, 12, ~~13~~, ~~21~~, ~~25~~

min = 3 Q_1 = 4 med = 9 Q_3 = 12 max = 25

By calc: 1var stats (L_1 = #s) → min = 3, Q_1 = 4, Med = 9, Q_3 = 12, max = 25



11.4 – The Normal Distribution

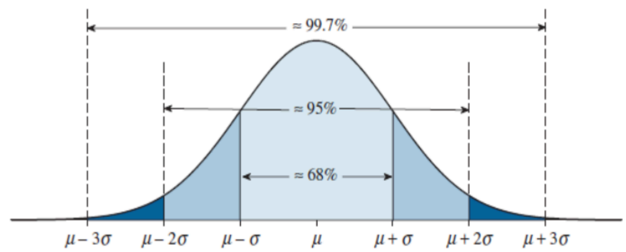
Empirical Rule (68 – 95 – 99.7 Rule)

"step"

68% of the data lies within 1 st dev of the mean.

95% of the data lies within 2 st devs of the mean.

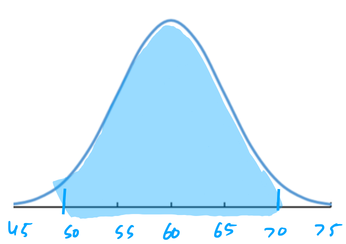
99.7% of the data lies within 3 st devs of the mean.



- Finding probabilities using the Empirical Rule.
  - Step 1 → **Draw** and **label** curve.
  - Step 2 → **Shade** curve.
  - Step 3 → **Use empirical rule**.

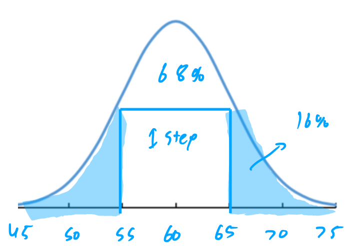
Example

Oak tree heights are normally distributed with mean 60 m and st dev 5 m.



a) Find the percent of trees between 50 m and 70 m tall.

2 steps ⇒ $\boxed{95\%}$

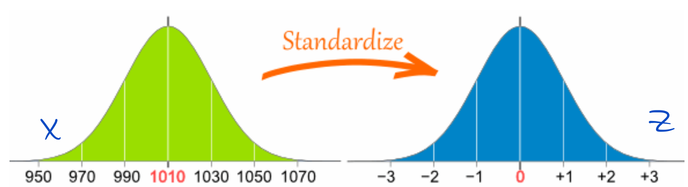


b) Find the percent of trees greater than 65 m

Outside = Total 100% − Inside 68% = 32%

ONLY right = $\frac{32\%}{2}$ = $\boxed{16\%}$

Finding probabilities based on the normal distribution



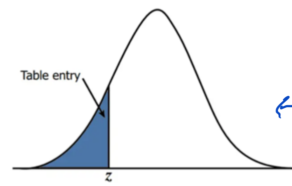
- Step 1 → **Standardize** using the **z-score**.

  Formula: $\boxed{z = \frac{x-\mu}{\sigma} = \frac{obs - mean}{st\ dev}}$

  - Ex) X has a normal distribution with mean 10 and st dev 2.
    Find the z-score for $X = 13$.

    $z = \frac{13-10}{2} = \boxed{1.5}$

- Step 2 → **Draw**, **label** and **shade** curve.
  - This is how you show your work!!!

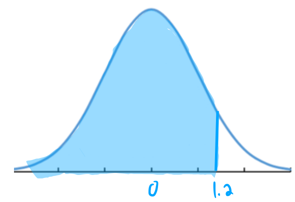


- Step 3 → Use '**Standard Normal Distribution'** table to find the probability for Z.
  - Table ALWAYS gives probability LESS THAN Z: P(Z < z).

  - Examples (How to use table)

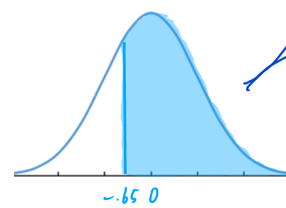
  - Left probability = TABLE (Directly)

    $P(Z < \underset{1.20}{1.20}) = \boxed{0.8849}$



  - Right probability = 1 - LEFT

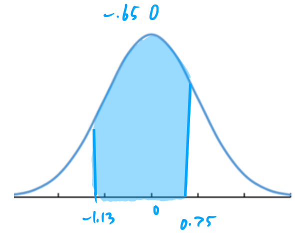
    $P(Z > -0.65) = 1 - P(Z < -0.65) = 1 - 0.2578$
    $= \boxed{0.7422}$



Does Not match
⇒ extra step

  - Between probability = LEFT Z_2 – LEFT Z_1

    $P(\underset{z_1}{-1.13} < Z < \underset{z_2}{0.75}) = P(Z < 0.75) - P(Z < -1.13)$
    $= 0.7734 - 0.1292$
    $= \boxed{0.6442}$

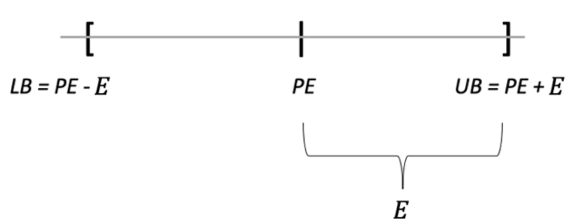


11.5 – Confidence Intervals

Point Estimates (PE)

- Using a statistic to estimate a parameter
  - Proportions: $\hat{p} = \frac{x}{n}$ and Means: $\bar{x}$

Margin of error

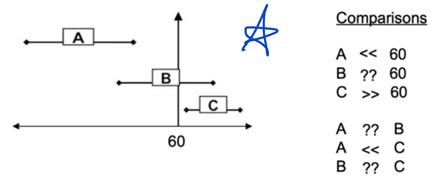


- C.I. = Point Estimate ± Margin of Error
  - E is the distance we extend our guess in both directions to form an interval

- Rule of thumb for margin of error in a survey
  - With 95% confidence, the margin of error, E, is approximately $\boxed{\frac{1}{\sqrt{n}} \cdot 100\%}$ for a sample of size n

- Interpretation (general structure)
    - I am C% confident that the true/population parameter + context is between (lower bound) and (upper bound).

- Comparing confidence intervals
    - When comparing confidence intervals to a particular value, or other intervals, we need to look at the ENTIRE interval to see if it is COMPLETELY below or above our comparison.



- Example: Out of 688 randomly selected students, 223 are members of at least one school club.
  a) Find the point estimate
  b) Find the lower and upper bounds of a 95% CI using the rule of thumb to calculate the margin of error.

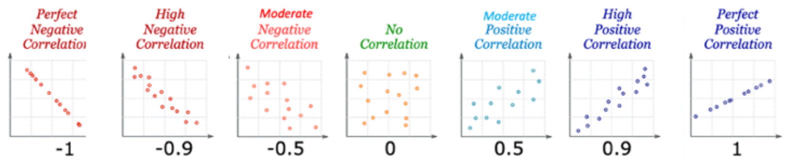
a) $\hat{p} = \frac{x}{n} = \frac{223}{688} \approx 0.324 \longrightarrow * 100\% = 32.4\%$

b) $E = \frac{1}{\sqrt{n}} * 100\% = \frac{1}{\sqrt{688}} * 100\% \approx 3.8\% \Rightarrow 95\%\ CI = \underset{PE}{32.4\%} \underset{\pm}{\pm} \underset{E}{3.8\%} = [28.6, 36.2]$

28.6 32.4 36.2
% at least one club

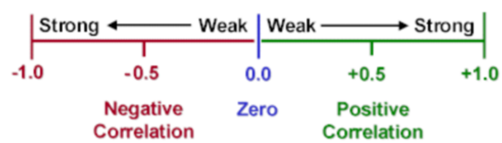
12.3 – Data Exploration

Scatterplots:



- **Form**: Linear, curved, or random scatter
- **Direction**: Positive, negative or no association
- **Strength**: Weak, moderate or strong

Correlation (r):



- Interpreting correlation (LINEAR)
    - Sign = Direction
    - Absolute value $|r|$ = Strength

- Calculate using calculator
    - **LinReg(ax+b) or 2-Var Stats** show work by writing this
    - $L_1 = X$, $L_2 = Y$

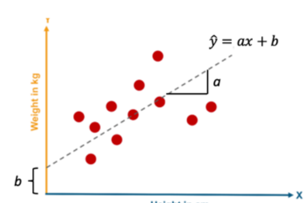
Regression:

- Step 1 → Determine if there is a **significant correlation (linear relationship)**.
    - Compare $|r|$ and Critical Value (CV) for n (sample size) and significance level α.
    - If $|r|$ > CV → statistically significant.

Critical Values of the Pearson Correlation Coefficient		
n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875

- Step 2 → Once we have a significant correlation, we can find the **regression line**.
    - $\hat{y} = ax + b$ (get results from correlation calculation)
      $= slope \cdot x + intercept$

- Step 3 → Make **predictions** using the regression line.
    - Just plug in the new X value to our equation and this will give us the predicted Y.



Example

Dataset:

X	3	5	4	7	6	10
Y	24	40	34	32	17	18

a) Calculate the correlation r.

2-var stats (x, y)

$\boxed{r = 0.4205}$

OR LinReg(ax+b) → $x = L_1$, $y = L_2$

b) Determine if r is significant for $\alpha = 0.01$.

$n = 6$

$|r| = 0.4205 < 0.917 = CV$

$\Rightarrow$ not significant

c) Suppose we have different regression equation where $\hat{y} = 5x + 2$.

Predict Y for $X = 3$:

$\hat{Y} = 5(3) + 2 = \boxed{17}$